*Evaluation is an information-gathering process, and it can unravel in much the same way as a detective story. **Just** as Sherlock Holmes redirects his crime-solving activities as new clues arise, evaluators **should be** committed to adapting their evaluation activities as new information is obtained.*

Decision-Oriented Approaches to Program Evaluation

Richard C. Larson
Edward H. Kaplan

Regardless of one's particular definition of evaluation, there can **be** no doubt that it is a process that produces information. This information may pertain to the extent **of** program compliance with contractual obligations, to managerial aspects of program operation, or to a test of some hypothesis about program functioning. Whatever the specific form **of** the information obtained from an evaluation, we would argue that the information is **useful** only ***to*** the extent that it informs **decisions**. For example, information that **exists** as data on a computer **tape is of** no value until and unless it is analyzed and used in some way. In just the same way, information not acted upon is no more valuable than no information at all.

The work of the first author **was** supported in part by grants nos. 78-91-AX-0007 and 80-IJ-CX-0048 from the National Institute of Justice of the **U.S.** Department of Justice to the Operations Research Center of the Massachusetts Institute of Technol***ogy.*** The **work** of **the second** author **was** supported in part by Public Systems **Evaluation,** Inc., 929 Massachusetts Avenue, Cambridge, Massachusetts.

Any decision can be viewed as an irrevocable allocation of resources. This definition of decision comes from Howard (1966), who explains that "irrevocable" implies not "for all time" but for the next short time interval that an allocation of at least one resource has been made. "That is, a decision is not a decision to make a decision but rather the concrete action implied by the decision. After any time interval, a decision may be replaced by another decision, perhaps as a result of updated information." Thus, if evaluation is a process that produces information useful in decision making, then evaluation is a process that produces information to assist in the allocation of resources. While this line of reasoning does not define the particular form that evaluation should take, we claim that it helps to clarify much of the debate over the various definitions of evaluation. Evaluation for compliance, evaluation for programmatic change, or evaluation for scientific inquiry—all can be reconciled by considering the decision makers who intend to use the information that comes from the evaluation.

The potential allocation of resources associated with an evaluation must be considered broadly. It can range from obvious programmatic changes (for example, shifting staffing patterns or operating procedures) to congressional authorization to spend more (or less) money on similar programs nationally or to a change in research efforts by one or more researchers who may want to devote more (or less) of their resources to that programmatic area. The allocation of resources can even refer to decisions by members of a program's client group to increase or decrease participation in the program.

Evaluation has a rich and diverse history. It seems rooted not only in substantive areas of concern, such as education, psychology, and mental health, but also in methodological areas of concern, such as classical statistics and experimental design. In the late 1960s and during the 1970s, the expanding role of government in providing services placed heavy demands on evaluation as a way of testing alternative ways for providing services. The demands on evaluation are not likely to diminish during the 1980s, but their nature may change as a result of increased interest in the accountability and productivity of service-providing agencies. Yet, these demands are being placed upon a field that amalgamates methodologically and substantively distinct areas of inquiry. Thus, there is a strong need for constructs and theories that serve to unify the heretofore disparate elements of evaluation. Evaluation for decision making is, we believe, one very useful framework in which to consider existing evaluation approaches and to identify the need for new approaches. It is our purpose in this chapter to review currently popular evaluation paradigms in their relation to decision making and to describe some new approaches that seem to contribute to the decision utility of evaluation.

We begin by reviewing the classical approaches of input, process, and outcome evaluation. Next, we will focus on three complementary quantitatively oriented methods for evaluation that can be used in conjunction with one or more of the classical approaches. We categorize these complementary methods as Bayesian approaches, adaptive methods of evaluation, and model-based methods. We will end with a brief summary and some conclusions.

Classical Approaches to Program Evaluation

Any operating public program is a process that acts on inputs to create outcomes. Many, perhaps most, classical evaluation paradigms for public programs have focused separately on program input, program process, or program outcome, so it is convenient to classify evaluations according to the program element being evaluated.

Evaluations that examine the resources or inputs dedicated to a given program are referred to as *input evaluations.* Perhaps the most limited type of evaluation, input evaluations are concerned with the inventory of personnel, facilities, and other resources assembled for the program; the proposed client or target group; and the program design. As a rule, primary attention is directed at compliance questions—for instance, whether program inputs as assembled and deployed comply with governmental or contractual regulations. One type of input evaluation is the program audit, which simply seeks to compare planned program expenditures for staff, facilities, and equipment with actual expenditures. Recently, such program audits have been extended into the area of program operations, and program managers have been required to fill out periodic reports of program activities and workloads, comparing and contrasting these with those planned in the approved contract or grant. Input evaluations are usually commissioned by program funding agencies or regulatory agencies; their findings are usually not directly useful to program personnel, except in cases where operations must change in order to achieve compliance. The decision makers who most often use the results of input evaluations are program funders, governmental regulatory agencies, and their equivalents.

While input evaluations focus on the resources directed to a given program, *process evaluations* examine the actual utilization of these resources by studying what the program physically does. A process evaluation of a program seeks to understand the causal mechanisms that translate the program inputs into program outputs or outcomes. If data about program outputs are unobtainable, a process evaluation tries to understand the mechanisms whereby program inputs are translated into action. Process evaluations use a mixture of qualitative and

quantitative techniques to understand the usually multifaceted nature of the program process. These include analysis of process-related performance measures; utilization of participant observers, interviews, and questionnaires; and other methods for understanding the total environment of the program. The outlook is Bayesian rather than statistical, meaning that impressionistic information can play a role which equals that of statistical information. Process evaluation can yield information that helps to establish the degree of program influence on observed outcomes. Process evaluations can also identify unintended side effects of the program. The findings of a process evaluation are often directly relevant to managers of the program being evaluated, but in many cases their usefulness to personnel and similar programs elsewhere is limited, owing largely to the idiosyncratic nature of the program and the specificity of evaluation conclusions.

What process evaluation does not attempt to determine is whether the program achieved its stated goals. This question is addressed by *outcome evaluations.* Several methodologies are available for performing outcome evaluation, but it is clear from the evaluation literature that experimental designs currently constitute the preferred approach. As Rossi and Wright (1977, p. 13) claim, "There is almost universal agreement among evaluation researchers that the randomized controlled experiment is the ideal model for evaluating the effectiveness of a public policy." Indeed, if it is possible to conduct a randomized experiment, the programs may be analyzed by the procedures of classical statistics. However, as the design of a test diverges from that of the classical experiment, the rationale for the use of statistical evaluation devices is weakened. In realistic evaluation situations, the requisite degree of rigidity required to apply these statistical methods with any confidence is rarely achieved.

A wide range of decision makers can be interested in the results of an outcome evaluation, including legislators considering the institutionalization of the experimental program across a wider jurisdiction, researchers who wish to test one or more intervention theories, program managers who wish to revise program activities to better achieve ultimate objectives, or program funders interested in learning whether the program is achieving its objectives.

It is ironic that the most widely used statistical method in outcome evaluations—the statistical test of a hypothesis—is itself not decision oriented. Most typically, one poses the null hypothesis that the program had no effect on one or more outcome variables and tests that hypothesis with an experimental or quasi-experimental design. It has by now been well documented in the evaluation literature that this procedure tends to be biased toward acceptance of the null hypothesis (or, shall we say, toward not rejecting the null hypothesis), and neither

acceptance nor rejection of the null hypothesis is especially useful in the decision-making context. In addition to asymmetries regarding the likelihood of the emergence of alternative hypotheses, the classical hypothesis test is incapable of including the costs of various types of errors (for example, of accepting the null hypothesis when it is false or of rejecting the null hypothesis when it is true). Such costs or errors are extremely important in a decision-making framework, yet classical statistics has difficulty in dealing with them. Thus, we have a situation in which outcome evaluation, which potentially affects the largest pool of decision makers, is supported by an evaluation paradigm that is largely inconsequential for decision making. In our view, this is one of the major gaps in current evaluation methodology.

Bayesian Approaches to Program Evaluation

As we have argued above, evaluations should be designed and conducted with their relevance for decision making in mind. In fact, the decision to evaluate or not to evaluate is itself an important allocation of resources. Thus, when one seeks to develop a unifying theory for evaluation, one most naturally gravitates toward methodologies that aid in the analysis of decisions. This leads us to the exciting field of Bayesian statistics and decision analysis (see, for example, Keeney and Raiffa, 1976; Thompson, 1975).

The use of the terms *Bayesian method, decision analysis,* and *utility theory* has often been confused and misinterpreted in the evaluation field. In fact, there are several different components to the application of these methods in evaluation. One approach to the use of decision analysis in evaluation is seen in the work of Edwards, Guttentag, and Snapper (1975). These authors proposed to use utility theory to establish numerical preferences on alternative outcomes of a program and its evaluation. By applying simple single-attribute linear additive utility theory to program evaluation, they attempted to derive recipe-type procedures for conducting what they call *decision theoretic* or *Bayesian* evaluations. Another distinctly different application is described by Thompson (1975), who considers the very decision to evaluate as problematic. His application of these methods considers whether or not to evaluate and, if the decision to evaluate is made, which evaluation design to select. He argues that the best performance measure for an evaluation is the expected policy improvement consequence of the evaluation design, minus the expected cost of implementing the design. Here the policy improvement and cost units must be compatible, a requirement which should not be taken lightly in the complex field of evaluation. The mathematical operation of calculating expectation by averaging over alternative outcomes must be done before the evaluation

is undertaken. This requires inclusion of subjective probabilities over outcomes as well as the utility of alternative outcomes. Thompson calls the policy improvement component of the averaged quantity the prior information value or PIV of an evaluation. Invoking decision analysis, Thompson argues that the PIV has the following properties: A decision maker should not pay more for an evaluation than its PIV, although expenditure of any amount less than the PIV is justified in the absence of alternative evaluations; among competing evaluations of equal cost, the evaluation with the highest PIV should be preferred; and among competing evaluations of different costs, the evaluation with the highest net PIV—the prior information value less the cost of the evaluation—should be preferred (Thompson, 1975, p. 40). The policy benefits underlying the PIV derive from the possibility that information provided by the evaluation will lead to program improvement.

While we think that the PIV is a useful concept, we believe that it may not be appropriate to advocate its use in actual evaluation settings at this stage in evaluation research. Estimation of the expected policy benefits of an evaluation and standardization of the benefits and costs is simply too difficult a task for all but the simplest situations. This limitation of the PIV concept does not, however, preclude its conceptual utility as a means of answering the question of whether to evaluate and of selecting the appropriate evaluation design.

Other Bayesian or decision-oriented methods relevant to evaluation focus on Bayesian statistical methods. Here, we refer specifically to Bayesian hypothesis testing and Bayesian methods of parameter estimation. One highly desirable attribute of Bayesian statistical methods is the ability of the Bayesian structure to include both impressionistic and hard information compatibly within one parameter estimation framework. Thus, for instance, an evaluator doing both a process and an outcome evaluation may have both process and relative frequency-type knowledge of one or more parameters; the process or impressionistic knowledge can be utilized to create a prior distribution on the unknown parameters, and the relative frequency information can be used to update the prior distribution to obtain a posterior distribution. While this approach may seem to be attractive in an evaluation setting, we are unaware that any completed evaluations have actually used this approach. Critics of the approach argue that the procedure is subject to large abuse, in the sense that prior beliefs and prejudices can be incorporated into a prior distribution under the guise of utilizing process information and bias the evaluation results as the evaluator desires. While this is certainly true, we would argue that the allegedly more scientific classical statistical hypothesis test is subject to equal abuse (for example, one could cleverly select a null hypothesis that would be difficult to reject for small and moderate sample sizes). The Bayesian parameter

estimation approach has been used to reanalyze evaluations of time series studies, such as the saturation patrol study reported by Schnelle and others, 1977; Willemain (1978a) has applied Bayesian procedures in a reanalysis of Schnelle's data.

One of the most relevant applications of Bayesian ideas in an evaluation framework is the Bayesian hypothesis test. The Bayesian hypothesis test is, in fact, less a hypothesis test than a procedure for deriving an optimum decision rule. Thus, decision making is its primary focus. In the simplest form of the Bayesian hypothesis test, there are two possible states of nature: state 0 and state 1; and two decisions that one can make are: to say that state 0 is the true state of nature or to say that state 1 is the true state of nature. The idea is to incorporate one's prior knowledge, the costs of various kinds of decisions and their concomitant errors, and recently acquired data into a unified framework for decision making.

Perhaps these ideas are best illustrated by a simple example from the medical field. Suppose we have a patient who is being tested for the presence of cancer; that is, the two states of nature are state 0, representing no cancer, and state 1, representing cancer. The decision to be made by a physician is whether or not to operate on the patient in the hope of eradicating the cancer. The risks are obvious: not to operate in the presence of cancer seems to carry the greatest risk; to operate in the absence of cancer yields unnecessary risk; not to operate in the absence of cancer is clearly the best possible situation. All of these can be summarized succinctly by four costs:

C_{00} = the cost of not operating in the absence of cancer
C_{01} = the cost of not operating in the presence of cancer
C_{10} = the cost of operating in the absence of cancer
C_{11} = the cost of operating in the presence of cancer

Our prior beliefs regarding this patient derive in part from our knowledge of the probabilities for patients from similar populations reviewed in the past. That is, we assign a value of P_0 to the probability that the patient in fact does not have cancer and P_1 to the probability that the patient in fact does have cancer. From past experiences, we have found that a fraction P_0 of similar patients did not have cancer and that a fraction P_1 of similar patients did have cancer.

The physician then carries out a test on the patient, and this test yields a numerical score; for example, the density of abnormal blood cells on a microscope slide. On the basis of this score, the physician will determine whether to operate. The second element of our prior belief is brought into play at this point. It has been found that patients who do not have cancer yield a distribution of score values centered around a

particular value S_0 of the score, while individuals who do have cancer yield another distribution of score values centered around S_1. Unfortunately (from the decision maker's viewpoint), the two distributions overlap, and this creates ambiguity in the score results and a potential for decision errors.

A decision rule, which is the outcome of the analysis of the Bayesian hypothesis test, is a partitioning of score values so that one has regions of score values in which the physician decides to operate and regions in which he decides not to operate. Any particular partitioning or decision rule has an associated error of deciding to operate when in fact there was no need, and an associated error of deciding not to operate when it was needed. Suppose for a particular decision rule that we call α the probability of deciding to operate when in fact the patient does not have cancer and β the probability of deciding not to operate when in fact the patient does have cancer. These α's and β's are the same α's and β's that one obtains in the classical statistical hypothesis test. Letting TC represent the total expected cost of a given decision rule,

$$TC = P_0[C_{00}(1-\alpha) + C_{10}\alpha] + P_1[C_{01}\beta + C_{11}(1-\beta)]$$

Minimization of the total expected cost is the emphasis of the analysis in a Bayesian hypothesis test.

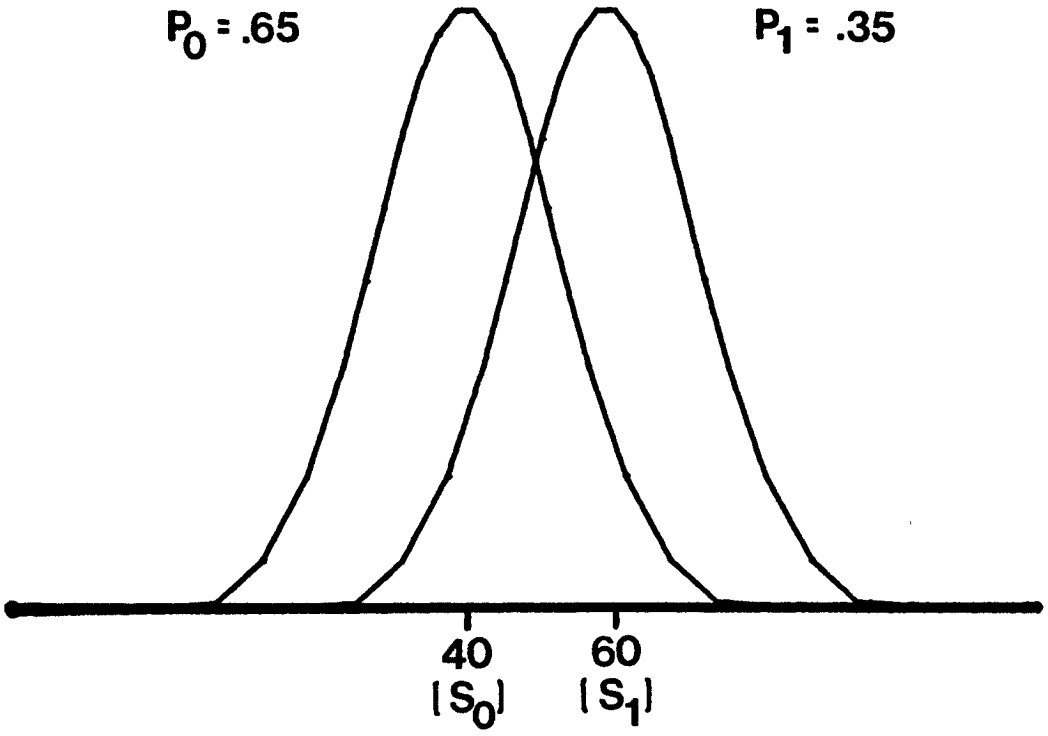
In this example, if the various costs involved relate to life expectancies, then the largest life expectancy would accrue to the patient who is not operated on and who does not have cancer, while the shortest life expectancy would accrue to the patient who does have cancer but who is not operated on. Since years of life expectancy are a positive good, to translate them into costs we must multiply by -1. As an example, suppose we take the following values for costs:

$$\begin{aligned} C_{00} &= -30 \text{ years} \\ C_{01} &= -\ 5 \text{ years} \\ C_{10} &= -25 \text{ years} \\ C_{11} &= -20 \text{ years} \end{aligned}$$

We also assume that $P_0 = 0.65$ and $P_1 = 1 - P_0 = 0.35$. Suppose further that the distribution of test scores for those who do not have cancer is normal with mean 40 and variance 169, while the distribution for those who do have cancer is normal with mean at 60 and variance at 169. The situation is diagrammed in Figure 1.

Utilizing all this information, it is not difficult to evaluate alternative decision rules (see Table 1). The optimal decision rule is a threshold score value equaling 45.95. For score values less than this amount,

Figure 1. Normal Frequency Curves of Test Scores for Patients With ($S_1 = 60$) and Without ($S_0 = 40$) Cancer



the physician decides to operate. This decision rule minimizes the total expected cost of the procedure, where cost, as we have indicated, is measured in years of life expectancy. We may contrast the life expectancy resulting from the optimal policy with that resulting from alternative, sometimes naive policies. The optimal decision rule results in a life expectancy of 24.71 years. However, if the physician is completely risk-averse and always decides to operate, then the life expectancy reduces to 23.25 years, but if he decides never to operate, the life expectancy is reduced to 21.25 years. A physician whose testing procedures left much to be desired might decide at random to operate on 35 percent of the patients; such a policy would result in a life expectancy of 21.95 years. In contrast, if one adhered strictly to α and β values found in the popular statistical tests, one would find that for an α of .05, there would be a corresponding β of .5424 and a life expectancy of 23.49 years; for a β of .05, there would be an α of .5424 and a life expectancy of 24.47 years. We see that blind adherence to any prespecified values for α or β will, in general, yield nonoptimal decisions. This same limitation on α values of .05 arises in many elements of outcome evaluation in which the evaluators attempt to apply statistical recipes to their evaluations which yield decisions that may be far from optimal.

Table 1. Life Expectancies Corresponding to Alternative Decision Rules

Decision Rule	α	β	*Life Expectancy*
1. Always Operate	1.0	0.0	23.25 Years
2. Never Operate	0.0	1.0	21.25 Years
3. Operate on 35 Percent of All Patients	0.35	0.65	21.95 Years
4. Fix $\alpha = .05$	0.05	0.5424	23.49 Years
5. Fix $\beta = .05$	0.5424	0.05	24.47 Years
6. Optimal Decision Rule (Minimize Total Cost)	0.3236	0.1399	24.71 Years

Adaptive Evaluation Design

It is our belief that the process of evaluating public programs should often be adaptive rather than fixed. One need for adaptiveness is clear: if the program being evaluated changes, evaluation activities may need to be adjusted, if they can be, to compensate for the change. Another is less obvious. Information gathered during the course of an evaluation arrives in a probabilistic manner, providing evaluators with different knowledge profiles about the program at each intermediate point during the evaluation. Some, perhaps most, knowledge profiles should suggest a change in evaluation activities. Armed with Bayesian decision-oriented concepts discussed in the previous section, one can develop a consistent structure for considering alternative evaluation designs.

To formalize these ideas, evaluation activities are directed by an evaluation design, which we define as a set of rules indicating how scarce evaluation resources are to be allocated over time and task. A fixed evaluation design is one in which the rules are inflexible or deterministic and do not allow for change at intermediate points during the evaluation. An adaptive or flexible evaluation design is one in which the rules are stated in a manner that is dependent on the evaluator's knowledge profile at intermediate points during the evaluation.

One difficulty in developing new methods of program evaluation is knowing when you have something that is in some sense better than what was used before. Thus, we need measures that allow us to compare the evaluative quality of the methods that we develop to the evaluative quality of more standard methods. This problem is not limited to researchers. Practicing evaluators, too, require valid and systematic means for comparing feasible evaluation designs in order to make an intelligent choice. (We do not use the word *optimal* here, because it is too strong for the problem setting.) To our knowledge, the idea of comparing alternative evaluation designs by means of an integrated set of performance measures attributable to the designs is new in the field of public program evaluation.

In a strict decision-theoretic sense, the best performance measure for an evaluation is the net PIV, as argued in the preceding section. Nevertheless, estimating the expected policy benefits of an evaluation and standardizing units of benefits and costs is too difficult a task for all but the simplest situation. Thus, at least in the short term, one should seek performance measures for evaluations that are easier to implement than the PIV. One class of promising evaluation design performance measures consists of measures that focus on particular evaluation performance characteristics, given that it has been decided to perform an evaluation with specified inputs. That is, one would emphasize selection of the preferred design, as indicated by one or more performance measures, within certain operational constraints involving time, money, personnel, or other resources devoted to the evaluation. This strategy is more compatible with evaluation practice, in which many if not most evaluations are commissioned by a Request for Proposal, a process that almost invariably places time and budget constraints on the evaluation.

We have just begun research on adaptive evaluation designs and we are exploring a number of evaluation design performance measures, including the information expected from an evaluation as measured by entropy or other indexes of information content, various measures of parameter variability as indicated by updated Bayesian density functions (for example, the variance of an updated or posterior distribution), and more traditional statistical measures obtained from classical statistics.

Equipped with a set of performance measures for evaluation designs, we plan to apply a number of them in particular evaluation problem situations. One example involves adaptively determining the various time periods of an evaluation. An evaluation is often partitioned into distinct time periods, with each period representing either a different set of evaluation activities or a different program implementation stage. During a three-period evaluation, for example, the evaluator might collect baseline data, during-program data, and post-program data. An important part of any evaluation design is the determination of the length of each period, usually with a constraint that limits the length of the entire evaluation and probably with certain other evaluation resource constraints. For instance, a very simple one-year, two-period fixed design would allot the first six months to the collection of baseline data and the second six months to the collection of new program data (assuming that a revised program is implemented precisely at the six-month mark). While most evaluation designs fix the lengths of the various evaluation time periods, circumstances arise in practice that make it possible adaptively and dynamically to determine the length of these periods. Basically, the purpose of each period is to col-

lect certain information within acceptable tolerances of uncertainty. Evaluation resources devoted to each period result in costs that vary linearly with the length of the period. Here, the problem of adaptive evaluation design is to determine dynamic decision rules for switching times between one evaluation period and the next, where accumulated costs and known information content are balanced against expected costs and information content of future decisions. The problem may even be extended from the adaptive determination of period durations within a given sequence to the optimal sequencing of evaluation periods.

Preliminary work by Willemain (1978b) and Willemain and Hartunian (in press) examines what is probably the simplest design issue for time series comparisons, namely, determining the division of a prospective time series comparison between simple baseline and experimental phases. This work assumes that the aim of experimental intervention is to change the rate of a Poisson process, as the approach grew out of reanalysis of a Nashville experiment with saturation-level patrols which established that the count of serious crimes can be well described by the Poisson probability law (Willemain, 1978a). Because most time series analyses have a retrospective baseline, the length of the baseline is usually determined by the availability of data. We found no guidance in the literature on how to determine the baseline duration prospectively when faced with an overall budget constraint. The analysis produced a method for optimal division of an evaluation budget between the two phases of the time series comparison. The optimal division depends on four key parameters: the budget level, the relative daily cost of the two phases of the experiment, the prior expectation of the baseline rate, and the prior expectation of the experimental impact. When the experiment is poorly funded, the choice of baseline duration becomes critical for obtaining the best possible estimate of the experimental impact, and proper design can greatly improve the efficiency of the evaluation. When this situation is considered to be adaptive, the duration of the baseline is made to depend on the baseline data as they appear. If the empirical results unfold in a way that confirms the prior estimate of the baseline rate, then the baseline has served its purpose and the resources can better be invested in the experimental phase of the study.

Optimal fixed and optimal adaptive designs were considered by Willemain and Hartunian (in press) in very simple two-period situations. Our current work is aimed at extending these analyses to more complex and realistic situations.

Another application of adaptive evaluation design that is responsive to information profiles as they appear at various times in an evaluation addresses adaptive allocation of evaluation resources. Viewed analytically by the methods and performance measures used to study the

previous problem, this problem appears to be quite similar. Substantively, it is quite different, and it is applicable to a far broader range of evaluation acitvities, particularly in process evaluation, where analytical guidance has been scant. Recalling that evaluation is an information-gathering process, it can unravel in much the same way as a detective story, and just as Sherlock Holmes redirects his crime-solving activities as new clues arise, evaluators should be committed to adapting their evaluation activities as new information is obtained. Our present work is aimed at such evaluation activities as adaptive allocation of interviewer and participant observers in process evalution. We anticipate that a report on this work will be available early in 1982.

One other major element of our adaptive evaluation design research focuses on the response of an evaluation to changes in the program being evaluated. We start with the premise that any program operating in a public agency is a dynamic process, with individuals maturing and changing their priorities, with operations evolving over time, and with new constraints imposed by regulatory agencies, new legislation, citizens groups, and so on. Any evaluation design that assumes a rigidly fixed world for any considerable amount of time is likely to experience difficulty as a result of these natural changes over time. Some of these processes have been recognized in the evaluation literature, such as maturation and attrition in Campbell and Stanley's (1963) now famous "threats to internal validity." One of the objectives of randomization is to structure an experiment so as to have these time-varying processes affect the control and experimental groups equally. In this way, measured differences between control and experimental outcomes cannot be attributed to these changes. The issue becomes much more problematic when randomization is not employed, as in time series quasi-experiments.

In our current research, we distinguish between incremental program changes and major changes, which we call fundamental changes. For the incremental type of change, we are exploring various methods and techniques that could be used to correct the evaluation data for the variability caused by incremental program change. Our primary method is based on modeling techniques, as discussed in the next section. For the fundamental type of change, we are exploring the possibility of incorporating the risk of that type of change a priori in the evaluation design process. For the incremental program changes, we believe that the soundest way to ensure the integrity of evaluation is to build simple mathematical models that correct for the gradual or sudden changes that arise in the environment; the assumptions in these models should be stated clearly and, when possible, subjected to empirical testing. An example will be described in the next section.

In our study of fundamental program changes, we note that no

evaluation design can stand impervious to all changes in the program or in its environment. One would like to incorporate the possibility of fundamental change in a program or its environment—change so great that the evaluation cannot continue—into the evaluation design process. Our point here is to catalogue the four or five likely causes of such catastrophic change and to assess the probability of each. One could use laws of combining probabilities to obtain a good Bayesian estimate of the probability of failure. If that probability exceeds some threshold, then the evaluation design as originally planned is unworkable. One may ask, "What is the alternative?" As an example, if the original evaluation were to occur at a single site, one clear alternative that appears to warrant serious consideration is multisite evaluation. Due to budget constraints, the cost of multisite evaluation is a smaller sample size at each site. However, the benefit is an increased likelihood that at least some useful information will be obtained by the end of the evaluation.

In general, in this part of our research on adaptive evaluation designs, we seek to identify the institutional, political, legal, and technical factors that create fundamental change in a program, environment, or both. We seek to determine the reliability with which the risks of such changes can be estimated. We seek also to develop probabilistic procedures, borrowing concepts from reliability theory and using relevant performance measures to develop less risky alternatives to single-site designs and other standard approaches.

Model-Based Methods for Evaluation

Closely linked to Bayesian and adaptive evaluations are the methods of model-based evaluation. A model is a conceptualization of a process or part of a process. It represents a hypothesized set of relationships that acts on the set of independent variables to produce values for the set of dependent variables. A model does not need to be mathematical; in fact, verbal or narrative models are often much richer than corresponding mathematical models. In evaluation, models offer aid for evaluation activities by systematically considering the set of causal linkages that acts on inputs to produce outcomes (that is, a model is a hypothesized set of relationships that, through program processes, transforms program inputs into program outcomes); by uncovering the importance of certain performance measures; by discovering the presence, cause, and magnitude of unintended biases in a data sample; by estimating the necessary duration of the evaluation at the design stage in order to assure that equilibrium, rather than transient program behavior, is observed and evaluated; in general, by building an evaluator's insight and intuition, thereby providing a systematic framework for generating and testing hypotheses and assisting in the exploration of

counterintuitive or unintended results; and by providing a basis for evaluating evaluations.

The use of modeling in an evaluation is quite compatible with the Bayesian and decision-oriented approaches discussed above. In fact, utilization of Bayesian statistical procedures requires probabilistic models as part of the Bayesian analysis. In a decision tree context, estimation of the probability with which certain events will occur is made easier by the construction and analysis of appropriate models at the appropriate branch of the decision tree. Our recommendation for the use of modeling in evaluation goes further than strict technical application of models; we argue that hypothesized or conjectured relationships, whether mathematical or verbal in nature, should be used throughout the evaluation enterprise to provide a coherent structure for carrying out that enterprise. For example, in process evaluation, without one or more conjectured models that link program inputs through process to program outcomes, one would find it extremely difficult to structure a plan for collecting information in one's evaluation activities; Carol Weiss calls such conjectured models for process evaluation "causal chain" models (Weiss, 1971, p. 140).

The recommendation for using models in evaluation is not new. For instance, Brenner and Carrow (1976) presented views very similar to ours. They argued that proper evaluation must specify the theoretical rationale whereby the criminal justice system ought to affect the crime rate or some other outcome and control for the effects of all significant variables that would of themselves significantly affect the crime rate or other outcomes.

"Recent methodological developments in the social sciences have pointed to the appropriateness of constructing a structural model by which causation is understood to flow from the interaction of a number of different variables. Such a model, resembling those in physical sciences, includes integrating different theories, all of which have a simultaneous effect on the outcome phenomenon" (Brenner and Carrow, 1976).

There are essentially two ways to generate models in evaluation activities: top-down or bottom-up. Top-down implies that underlying theories or hypotheses give rise to the models. Bottom-up implies that the entire universe of models has been considered and that one or more of these are statistically discovered from the evaluation data. There are numerous problems with bottom-up modeling, which in application resembles little more than curve fitting. The flow of causation is hard to determine, and such things as correlations with invisible third variables can confound and confuse results. Our preference is for utilization of top-down or deductive models whenever possible.

Deductive models do not have to be complicated to provide insights. Here are three examples to illustrate this fact:

Institutional Half-Life. Suppose that we are considering the evaluation of a major reform or intervention in a public institution that has a twenty-five year retirement plan for its employees. Suppose further that the institution is in an equilibrium mode with regard to the number of employees; that is, it is neither growing nor diminishing in number of employees. Many major interventions are known to be impeded by institutional memories, which diminish only as the older employees leave to be replaced by new, younger employees. An evaluation designer could argue, then, that any major intervention in such an institution would only reach equilibrium or steady state once a good fraction of the sources of institutional memories has left the agency. One measure of the duration required for such an exodus is the institutional half-life of the institution, which is defined as the time required for a 50 percent turnover of personnel. In the institution just discussed, with a twenty-five-year retirement plan, the institutional half-life would be perhaps five or ten years if most employees stayed until retirement. However, this amount of time is quite sobering to a potential evaluator, who may want to consider an evaluation over a six-month, one-year, or even two-year period. Indeed, the idea of institutional half-life may suggest to the evaluator that any evaluation that is feasible within the short term must of necessity occur shortly after the intervention.

Crime and Streetlighting. Sometimes, simple modeling concepts will suggest alternative interpretations of evaluation performance measures that can change the appearance of failure into an appearance of success or vice versa. As an example, consider a city that is implementing a new strategy for combating street crime along one major street in the city. Suppose that, before the program, the number of street crimes per month along that major street was ten, and suppose that following implementation of the new program, the number of street crimes per month increased to twenty. An evaluator of this program could conclude that the program had been a failure, since the number of crimes doubled following the program. However, these data could be reinterpreted by considering the probability of an individual pedestrian being victimized while on the street. If the pedestrian volume following the new program quadrupled, then the *per-pedestrian* victimization probability would be half what it was before the program. In this light, the program appears successful. We are not arguing here that a single interpretation is correct but that the incorporation of one simple modeling idea—the flow of pedestrians and pedestrian-specific victimization probability—can sharply alter one's view of the success or failure of a program.

Interbeat Dispatch Frequency. Within police circles, a police officer is usually required to respond to as many calls for service within his or her own beat as is physically possible. That is, police administra-

tors require that an officer assigned to a given beat handle as large a fraction of the calls for assistance on that beat as possible. This enhances officer identity with the beat, and it also enhances citizen satisfaction, since citizens get to know the officers assigned to their beat. However, any evaluator evaluating a program involving police patrol forces should be aware of a physical law that limits the dispatch assignments within an officer's patrolling beat. The result of this law is as follows: On the average, if patrol officers in an area are busy X percent of the time, then approximately X percent of patrol dispatch assignments are to officers in beats other than that in which the call for service originated. That is, for officers busy X percent of the time, roughly X percent of their dispatches take them to beats other than their primary beat. Also, citizens see officers responding to their calls for service from other beats roughly X percent of the time. This result is derived from a simple modeling concept. Since calls for service arrive in a completely random and unscheduled manner, if a patrol officer is busy X percent of the time, then X percent of the time a call for service that arrives from his beat will find him already busy. Thus, a dispatcher who requires a rapid response to that call for service must find and dispatch an available unit for that X percent of calls in a contiguous or nearby beat. What is true for one beat is true for all. The law can be modified only by altering the dispatcher's procedures, such as deliberately delaying certain calls for service and entering calls in queue, in order to await the availability of the beat car.

Very simple models such as those illustrated above could greatly enhance the evaluator's insight and intuition into program process and program outcome, and this could lead to improved evaluation designs, executions, and analyses.

As we have demonstrated, informal, back-of-the-envelope models can often be employed in evaluating public programs. There are situations, however, where the degree of program complexity warrants the use of more sophisticated models. To model such large-scale programs can require the aid of a computer.

Computer-based models have recently been used in evaluation of large-scale housing and criminal justice programs. For example, researchers at the Urban Institute developed a detailed simulation model of the urban housing market; this model has been employed to analyze various housing programs, including the HUD-sponsored housing allowance experiments (Carlton and Ferreira, 1977; de Leeuw and Struyk, 1977). This model was used to assess the implications of alternative housing assistance payment formulas, so that the most desirable subsidy scheme could be identified. Also, the model can determine the fraction of income spent on housing under alternative assumptions governing the elasticity of demand for housing; this is obviously quite

important in the analysis of housing allowance programs. The computer-based JUSSIM (Justice Simulation) model has been used to examine the entire criminal justice systems of a number of states (Blumstein and Larson, 1972). JUSSIM enables one to trace entire cohorts of offenders through the criminal justice system. Thus, JUSSIM aids in studying court and prison congestion, the effects of sentencing on the size of the prison population, and the timing of various levels of recidivism (for example, time until rearrest versus time until reconviction). Although the models referenced here are quite complex, the rationale for using them is the same as the rationale for using simple models. In the examples that we have cited, researchers were able to examine the likely outcomes of alternative programs in a systematic manner.

Summary and Conclusions

Throughout this chapter, we have argued for the development and utilization of a coherent, decision-oriented approach to evaluation. Many details of what we have proposed have yet to be worked out in complex evaluation settings, and we would be naive to claim that everything we propose could be implemented without difficulty. Yet, it is our contention that there is a clear need for one or more coherent and comprehensive approaches to this amalgam of heretofore disparate fields. We have chosen a decision-oriented approach because we believe that the primary purpose of single-project evaluations is to inform one or more decision makers.

Much of what we have proposed involves formal methods, requiring mathematical modeling, Bayesian statistics, and so on. The literature on operations research alone yields such relevant tools of analysis as probabilistic modeling, decision analysis, Bayesian analysis, and Markov decision processes. Other relevant tools come from the fields of optimal control theory, information theory, and stochastic processes. While these methodologies are indeed powerful, their most successful applications have tended to be in highly technical areas such as process control in chemical plants, oil exploration, space exploration, and operation of communications systems. Only recently have we begun to see successful limited applications of these techniques to more human-oriented areas such as the forecasting of personal career trajectories, the location of human services facilities in cities, the allocation of municipal services resources, the prediction of recidivism patterns, and even the identification of sources of selection bias in criminal justice evaluations.

Analytical techniques such as these work best in precisely structured environments with readily measurable input, process, and outcome variables. They tend to become analytically intractable when imbedded in an imprecisely formulated environment. These attributes

would appear to limit the applicability of modern quantitative techniques within the imprecise field of evaluation. We believe that this is a correct conclusion if one judges the results of using these techniques by the infallible "recipes" that they generate for the naive evaluator. However, we believe that a second yardstick is required. We do not propose using these techniques as foolproof recipes in evaluations. Rather, we believe that the careful and judicious use of abstractions in evaluation settings can yield valuable insights to the evaluation designer and implementer that greatly improve one's understanding of the many facets of evaluation. Armed with improved insight and intuition, the evaluation designer will have analytically grounded guidance as to the consequences of alternative evaluation designs that he or she may consider.

The situation is somewhat analogous to that of business school students who are taught decision analysis and Bayesian statistics for use in complex and ill-defined business settings. By working through the analysis of abstractions of business problems, the potential decision maker gains insight into the dynamic, adaptive decision situations that will confront him or her every day as a manager. Yet, rarely as a manager will he or she in fact sit down to diagram a decision tree. The process has to be integrated into one's intuitive decision-making process.

By applying these techniques in an evaluation setting, we hope to develop similar useful abstractions and analyses for evaluators. The alternative is to ignore the need for comprehensive conceptualizations of the evaluation enterprise. Analytically, it is certainly much easier to borrow from other analytical areas, such as classical statistics, when necessary. But to allow the increased difficulty to be the determining argument against decision-oriented Comprehensive evaluations would be unfortunate. Decision-oriented approaches show considerable promise for reducing the misallocation of expensive evaluation resources, the collection of redundant information, haphazard responses to unexpected changes in the program, rote performance of statistical analyses without sequentially formulating and testing hypotheses, analysis and display of information without regard to the decisions to be influenced by it, and, most troublesome, the tendency to view rigid experimental design as the ultimate paradigm for evaluation.

References

Blumstein, A., and Larson, R. C. "Analysis of a Total Criminal System." In A. W. Drake, R. L. Keeney, and P. M. Morse (Eds.), *Analysis of Public Systems.* Cambridge, Mass.: M.I.T. Press, 1972.

Brenner, M. P. H., and Carrow, D. "Evaluation Research with Hard Data." In *Criminal Justice Evaluation.* New York: United Nations, 1976.

Campbell, D. T., and Stanley, J. C. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally, 1963.

Carlton, D. W., and Ferreira, J., Jr. "Selecting Subsidy Strategies for Housing Allowance Programs." *Journal* of Urban Economics, 1977, *4*, 221-247.

deLeeuw, F., and Struyk, R. J. "Analyzing Housing Policies with the Urban Institute Housing Models." In G. K. Ingram (Ed.), Residential Location and Urban Housing *Markets*. Cambridge, Mass.: Ballinger, 1977.

Edwards, W., Guttentag, M., and Snapper, K. "A Decision-Theoretic Approach to Evaluation Research." In E. L. Struening and M. Guttentag (Eds.), *Handbook of Evaluation* Research. Beverly Hills, Calif.: Sage, 1975.

Howard, R. A. "Decision Analysis: Applied Decision Theory." In D. B. Hertz and J. Melese (Eds.), *Proceedings of the Fourth International Conference* on Operational Research. New York: Wiley, 1966.

Keeney, R. L., and Raiffa, H. Decisions *with* Multiple Objectives: Preferences and Value *Tradeoffs*. New York: Wiley, 1976.

Rossi, P. H., and Wright, S. R. "Evaluation Research: An Assessment of Theory, Practice, and Politics." Evaluation *Quarterly*, 1977, *1* (1), 5-52.

Schnelle, J. F., Kirchner, R. E., Jr., Casey, J. D., Uselton, P. H., Jr., and McNees, M. P. "Patrol Evaluation Research: A Multiple Baseline Analysis of Saturation Police Patrolling During Day and Night Hours." *Journal of Applied* Behavior *Analysis*, 1977, *10*, 33-40.

Thompson, M. S. Evaluations for Recisions in Social Programmes. Lexington, Mass.: Heath, 1975.

Weiss, C. H. "Utilization in Evaluation." In F. G. Caro (Ed.), Readings in Evaluation Research. New York: Russell Sage Foundation, 1971.

Willemain, T. R. "Bayesian Analysis of Crime Rate Changes in Before-After Experiments." Report OR 0-75-78. Cambridge, Mass.: Operations Research Center, Massachusetts Institute of Technology, 1978a.

Willernain, T. R. "Analysis of a Contingent Experimental Design: A Before and After Experiment with a Baseline Period of Random Duration." Report OR 79-78. Cambridge, Mass.: Operations Research Center, Massachusetts Institute of Technology, 1978b.

Willemain, T. R., and Hartunian, N. F. "The Design of Time Series Comparisons Under Budget Constraints," in press.

Richard C. Larson is professor of electrical engineering and urban studies at M.I.T., where he is codirector of the M.I.T. Operations Research Center. He is also founder and president of Public Systems Evaluation, Inc., a nonprofit applied research firm.

Edward H. Kaplan is a doctoral degree candidate in the Department of Urban Studies and Planning at M.I.T.